DOCUMENT RESUME

ED 383 714 TM 023 135

AUTHOR Eignor, Daniel R.; And Others
TITLE The Effects on Observed- and True-Score Equating Procedures of Matching on a Fallible Criterion: A Simulation with Test Variation.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-90-25
PUB DATE Oct 95
NOTE 30p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Criteria; *Equated Scores; Item Response Theory; *Raw Scores; Sampling; Simulation; *Test Format; *True Scores
IDENTIFIERS *Equipercentile Equating; Levine Equating Method; Three Parameter Model; *Tucker Common Item Equating Method

ABSTRACT

Two recent simulation studies were conducted to aid in the diagnosis and interpretation of equating differences found between random and matched (nonrandom) samples for four commonly used equating procedures: (1) Tucker; (2) Levine equally reliable; (3) Chained equipercentile observed-score; and (4) three-parameter, item response theory true-score equating. For these simulations logistic, test forms were equated to themselves, a situation that does not pattern reality. In the current simulation, test variation was added as an additional variable for study. The results of the current simulation confirmed the results of the previous two simulations and support the prediction based on theoretical grounds that observed-score equating methods, such as Tucker and Chained equipercentile, are more affected by sample variation than are a true-score method (IRT) or an observed score method based on true-score assumptions (Levine equally reliable). The results further suggest that matching equating samples on the basis of a fallible measure of ability, such as anchor test score, is not advisable for any equating method studied except possibly the Tucker method. Three figures and one table present simulation results, and an appendix contains two additional tables. (Contains nine references.) (Author)

RR-90-25

RESEARCH REPORT

THE EFFECTS ON OBSERVED- AND TRUE-SCORE EQUATING PROCEDURES OF MATCHING ON A FALLIBLE CRITERION: A SIMULATION WITH TEST VARIATION

Daniel R. Eignor
Martha L. Stocking
Linda L. Cook

Educational Testing Service
Princeton, New Jersey
October 1990

The Effects on Observed- and True-Score Equating Procedures of Matching on a Fallible Criterion: A Simulation with Test Variation[1]

Daniel R. Eignor[2]
Martha L. Stocking
Linda L. Cook

Educational Testing Service
Princeton, New Jersey

Running head: Test Variation

---

[1]This study was supported by Educational Testing Service through Program Research Planning Council funding.

[2]The authors would like to recognize Maxine Kingston, Nancy Wright, and Miriam Feigenbaum for programming and data preparation assistance, Charles Lewis, Robert Mislevy, Ledyard Tucker, Neil Dorans, Marilyn Wingersky, and Ida Lawrence for psychometric advice, and Georgiana Thurston for typing the paper.

Abstract

Two recent simulation studies were conducted to aid in the diagnosis and interpretation of equating differences found between random and matched (nonrandom) samples for four commonly used equating procedures: Tucker, Levine equally reliable, and Chained equipercentile observed-score procedures and the 3PL IRT true-score equating procedure. For these simulations, test forms were equated to themselves, a situation that does not pattern reality. In the current simulation, test variation was added as an additional variable for study. The results of the current simulation confirmed the results of the previous two simulations and support the prediction based on theoretical grounds that observed-score equating methods, such as Tucker and Chained equipercentile, are more affected by sample variation than are a true-score method (IRT) or an observed-score method based on true-score assumptions (Levine equally reliable). The results further suggest that matching equating samples on the basis of a fallible measure of ability, such as anchor test score, is not advisable for any equating method studied except possibly the Tucker method.

The Effects on Observed- and True-Score Equating Procedures of Matching on a Fallible Criterion: A Simulation with Test Variation

INTRODUCTION

Recently, in an attempt to circumvent differences in common item or anchor test equating results across methods that are caused by samples that differ in ability, researchers at Educational Testing Service have begun to study the effects on the commonly used equating procedures of matching one of the equating samples being used to the other, through scores on the set of common items or anchor test. Lawrence and Dorans (1990) were the first to study the effects of matching, using anchor test scores on the Scholastic Aptitude Test (SAT), and a number of other studies of matching followed the Lawrence and Dorans work. These studies have recently been published as a set in an edition of Applied Measurement in Education (1990).

The work described in this paper may be viewed as an extension of the real data SAT matching study conducted by Lawrence and Dorans (1990) and two simulation studies involving matching with SAT data, one by Stocking, Eignor, and Cook (1988) and the other by Eignor, Stocking, and Cook (1990). Because of this, some of the details of the standard SAT data collection design will first be reviewed; then the results of the above-mentioned studies will be briefly discussed. For in-depth details about matching procedures, the reader should consult Dorans (1990).

Figure 1 depicts the basic SAT equating data collection design, which essentially represents an equating design linking the new form, labelled NEW, to two old forms OLD1 and OLD2. The specific old forms to be used in the

equating are established in the SAT braiding plan (Angoff, 1971); in general, the populations taking forms NEW and OLD1 will be populations of similar ability (data for form OLD1 will have been collected at a corresponding administration (same time of year) in a year previous to that in which form NEW was given), while the group of examinees taking form OLD2 will represent either a more or less able candidate population (data for form OLD2 will have been collected at a noncorresponding administration in a year previous to that in which form NEW was given). Form NEW is linked to OLD1 via one anchor test (EQ1) and to OLD2 via another anchor test (EQ2). These anchor equatings are performed using representative (random) samples from the populations taking the forms. Typically, the average of the anchor equatings to the two old forms is taken as the operational conversion for the new form.

Insert Figure 1 about here

In the Lawrence and Dorans (1990) study, the authors focused on the equating of NEW to OLD2 and performed both conventional observed-score (Tucker, Levine equally reliable, and Chained equipercentile; see Angoff, 1984, Chained equipercentile is Design V)) and three-parameter logistic (3PL) item response theory (IRT) true-score (see Lord, 1980) equatings under both representative (random) and matched (nonrandom) sampling conditions. In the matched (nonrandom) condition, scores on EQ2 were used in an attempt to match the ability level of the sample taking OLD2 to the ability level of the sample taking NEW. That is, the distribution of anchor test scores was made to be the same in the OLD2 and NEW samples, in the process altering the

characteristics of the OLD2 sample so that it was no longer a representative (random) sample.

Consistency of equating results, and particularly scaled score means, across the representative and matched sample conditions was used as the criterion in the Lawrence and Dorans study. Lawrence and Dorans found that the means for Tucker equatings varied the least across the two sampling conditions and that the means for the Levine equally reliable, Chained equipercentile, and 3PL IRT equating methods, while varying across the two conditions, also tended to converge to the mean from the Tucker equating under matched sampling conditions.

One potential problem with using consistency as the criterion is that consistent equating results may be different from the "true" equating results, were they known. In other words, the consistent Tucker equating results may have differed more from the "true" equating results in the Lawrence and Dorans study than the inconsistent Levine or IRT equatings. The lack of availability of "true" equating results suggested the need for a simulation study.

Stocking, Eignor, and Cook (1988) developed a general simulation model and then performed a sequence of simulations and subsequent equatings based on that model that addressed a number of specific issues in the application of both conventional (Tucker, Levine equally reliable, and Chained equipercentile) and IRT-based (3PL true-score) equating methodologies, many of which were brought out in the Lawrence and Dorans (1990) study. The purpose of their study was to investigate the impact on the four equating procedures just mentioned of: 1) differences in abilities of samples used for equating, both when each examinee has complete data (an unrealistic setting) and when

each examinee has missing data (a more realistic setting); 2) subsequent matching of samples on IRT ability, an infallible measure of ability (an unrealistic setting); and 3) subsequent matching of samples on anchor test observed score a fallible measure of ability (a more realistic setting). To be consistent with the Lawrence and Dorans study, the effect on scaled score means of these various experimental conditions was chosen for study.

The results of the Stocking et al. study the prediction based on theoretical grounds that observed-score equating methods, such as Tucker and Chained equipercentile, are more affected by sample variation than are a true-score equating method (3PL IRT) and an observed-score method based on true-score assumptions (Levine equally reliable). Their results further suggested that matching equating samples on the basis of a fallible measure of ability is not advisable for any equating method studied other than Tucker.

The results of the Stocking et al. study, i.e., the scaled score means and standard deviations, were not completely inconsistent with the Lawrence and Dorans (1990) findings for SAT-Verbal in that the Stocking et al. results corresponded fairly closely to the results found by Lawrence and Dorans for one of the eight verbal forms they studied. However, the Stocking et al. results were fairly inconsistent with results for the other verbal forms studied by Lawrence and Dorans and quite inconsistent with the Lawrence and Dorans findings for SAT-Mathematical. The conclusions of the Stocking et al. study were based on a single sequence of simulations, and because the results differed a good deal from the Lawrence and Dorans real data results, a replication of the Stocking et al. study was undertaken, using a different SAT-Verbal form and completely new samples. In addition, the samples used for

the replication were based on an examinee group that was a good deal more able (about a fifth of a standard deviation on the SAT scaled score metric) than the examinee group used to define samples in the original study. The results of the replication were reported in Eignor, Stocking, and Cook (1990). The results of the replication phase essentially confirmed the results of the Stocking et al. (1988) study and, collectively, the results from both studies provided a reasonably strong basis for making recommendations about whether to match on a fallible criterion, such as anchor test score.

However, in both of these simulation studies, the design called for variations in sample ability and the completeness of response data while controlling for test variation. Hence, tests were equated to themselves. While the results of the studies were seen by some as being informative, they do not pattern reality in equating the SAT, where a new form is equated to different old forms.

The purpose of the present study was to introduce test variation into the simulation procedure, thereby providing an indication of the effects of test variation over and above the effects of variations in sample ability and completeness of response data on the anchor test matching process. Outside of the introduction of test variation (i.e., there were three distinct forms being used in the equating, rather than one), all other elements of this simulation completely paralleled the previous two simulations (Stocking et al., 1988; Eignor et al., 1990). Selected results from the previous two simulations will be presented in this paper so they may be contrasted to the results obtained with the introduction of test variation.

THE STUDY DESIGN

The Definition of True Item and Person Parameters

For the sequence of simulations performed, true item and person parameters were required. They could, of course, have been invented. It was more realistic, however, to use existing parameter estimates, but treat them as if they were true. It seems reasonable to assume that such a definition of truth captures at least some of the predominant features of actual data, such as the spread of abilities and item difficulties. For this purpose, the results of a LOGIST calibration (Wingersky, Barton, & Lord, 1982) of an 85-item SAT-Verbal test form (administered in two separately-timed sections) plus a 45-item anchor test or equating section were used as the true item parameters for the new form (NEW) and equating section EQ1 (see Figure 1). The results of another LOGIST calibration of the same 85-item form plus a 40-item anchor test section supplied the true item parameters for EQ2. The results of a LOGIST calibration of another 85-item Verbal test form plus the associated 45-item anchor test supplied the true item parameters for OLD1. Finally, the results of a fourth LOGIST calibration of still another 85-item Verbal test form plus the associated 40-item anchor test supplied the true parameters for OLD2. All parameters were placed on a common scale using the characteristic curve transformation method (Stocking and Lord, 1983), applied to either the 45-item or 40-item anchor test from the separate calibrations. Forms OLD1 and OLD2 were the actual Verbal forms to which NEW was equated at its first operational administration.

True person parameters were defined to be the ability estimates obtained from a random sample of 3004 real examinees drawn from the total group that took NEW and its associated equating section EQ2. This total group had an SAT scaled score mean of 441 and scaled score standard deviation of 108.

Two population distributions of true ability were then defined. The first was defined to be exactly like the distribution of true person parameters, with mean true ability of -.02 and standard deviation of true ability equal to 1.05. A second population was defined to be less able, with mean true ability of -.35, but having the same standard deviation as the first population (1.05).

A total of seven independent samples of size N=3000 were drawn as follows:

Sample	Drawn from Population	Sample Mean Ability	Sample S.D. of Ability	To be Administered
1	1	-.03	1.05	NEW + EQ1
2	1	.00	1.07	NEW + EQ2
3	1	-.05	1.06	OLD1 + EQ1
4	1	-.03	1.05	OLD2 + EQ2
5	2	-.34	1.07	OLD2 + EQ2
6[1]	2	-.06	1.06	OLD2 + EQ2
7[1]	2	-.05	1.04	OLD2 + EQ2

The Generation of Response Data

Two types of response data were generated for each simulated examinee (simulee) -- complete data response strings and response strings reflecting missing data. Complete data response strings were generated in the standard

[1]Sample 6 was matched to sample 2 using the complete data observed formula-score distribution of sample 2 on EQ2. Sample 7 was matched to sample 2 using the missing data observed formula-score distribution of sample 2 on EQ2.

fashion using the simulee's true ability and the item's true 3PL parameters to generate the model predicted probability of a correct response, which was than compared to a random number selected from a uniform [0,1] distribution (see Lord, 1980).

The missing data response strings were generated from empirically-based models of speededness (for not reached items) and omitting behavior. With these models, both the number of items reached and the number of items omitted are functions of ability level. These models and the procedure for simulating the two kinds of missing data are described in detail in Stocking et al. (1988).

The Design of the Calibrations and Equatings

The simulated responses from the seven samples of simulees to the test forms and equating sections were combined into six separate concurrent LOGIST runs, each representing an experimental condition. The design of each LOGIST run was the same, and patterns the usual SAT data collection design presented in Figure 1:

	NEW	EQ1	EQ2	OLD1	OLD2
Sample 1	X	X			
Sample 2	X		X		
Sample 3		X		X	
Sample Y			X		X
(Y=4,5,6, or 7)					

The data for all samples in a LOGIST run were either complete or contained missing data. Sample 1 was administered the new form (NEW) and one anchor test (EQ1); Sample 2 was administered the new form (NEW) and another anchor test (EQ2); Sample 3 was administered the first anchor test (EQ1) and the

first old form (OLD1); and a final sample (either sample 4, 5, 6, or 7) was administered the second anchor test (EQ2) and the other old form (OLD2). The samples taking EQ2 and OLD2 in each LOGIST run, samples 4-7, were drawn in the following fashion. Sample 4 was drawn randomly from the same population as the other samples; Sample 5 was drawn randomly from the lower ability population; Sample 6 was drawn from the lower ability population to match the distribution of complete data observed formula-scores obtained by sample 2 on EQ2; and Sample 7 was drawn from the lower ability population to match the distribution of missing data observed formula-scores obtained by sample 2 on EQ2.

From the item parameter estimates derived from each of the LOGIST runs or from the observed-score data for the samples used in the runs, the new form was equated to each old form using the Tucker, Levine equally reliable, Chained equipercentile, and 3PL IRT equating methods. The two equatings were also averaged to produce a final equating. All old forms were placed on the SAT 200 to 800 scaled score metric by the nonlinear equating originally derived for each of the forms when they were given operationally for the first time as new SAT forms. Projected scaled score means and standard deviations were computed for each single equating and each average using samples of over 90,000 examinees who took NEW at its initial equating administration.

The Experimental Conditions

The series of simulations were designed to study six experimental conditions, shown in the following table, which contains a letter for each experimental condition:

	True Ability Distribution		
	Equivalent	Unequal	Equivalent by Matching
Complete data	A	B	C
Missing data	D	E	F

Condition A, complete data and equivalent samples, is a benchmark condition in that, while unlikely to be realized in practice, it represents the best circumstances for any equating method. In addition, samples have been chosen to be equivalent on the basis of an infallible criterion. Condition B, complete data and unequal samples, provides for the exploration of the effects of different sample abilities while still maintaining the ideal situation of complete data for all simulees. Condition C, complete data and matched samples, provides for the explanation of the effects of matching on a fallible criterion while still maintaining the ideal situation of complete data for all simulees. Condition D, missing data and equivalent samples, is a more realistic condition in that samples now incorporate missing data. In this condition, as in Condition A, samples have been chosen to be equivalent on the basis of an infallible criterion. Condition E, missing data and unequal samples, represents what is typically obtained in an SAT equating of NEW to OLD2 in the absence of any further data manipulation. Condition F, missing data and matched samples, represents the matching procedure employed by Lawrence and Dorans (1990); that is, matching samples on the basis of a fallible criterion in an attempt to achieve the ideal condition of equivalent samples.

RESULTS AND DISCUSSION

Table 1 shows the projected scaled score means and standard deviations for all individual equatings performed and for the averages. Tables A1 and A2 of the Appendix display comparable data from Stocking et al. (1988) and Eignor et al. (1990). In Figure 2, plots of the projected scaled score means for the individual equatings (not the averages) are displayed. Figure 3 contains comparable plots of the results from the Stocking et al. (1988) and Eignor et al. (1990) studies. In both Figures 2 and 3, the left side gives the results of the equatings of NEW to OLD1, and the right side gives the results for the equatings of NEW to OLD2. The experimental conditions are positioned along the horizontal axis. The projected scaled score means are read from the vertical axis. The points for a particular equating method are connected by dashed or solid lines, identified in the legend for each method, for the complete data cases and again for the missing data cases, to make the plots easier to read.

Insert Table 1 and Figures 2 and 3 about here

Table 1 and Figure 2 show that the differences among projected scaled score means are relatively small, although generally larger than the differences seen in Figure 3, where tests were equated to themselves. The importance of these differences among scaled score means is not possible to judge, however, since approximate standard errors of equating have not been

developed for all methods (i.e., the IRT standard errors of equating have not been developed to date).

To evaluate these results, it seems useful to compare the results of each equating method across experimental conditions to its own value in the "benchmark" condition. This condition, shown to the far left of each subplot, is the one in which data are complete for each simulee and all samples of simulees are drawn from the same ability distribution. In addition, this condition, along with the comparable missing data condition (condition D), represent "true" conditions in the sense that, in both cases, samples have been matched on the basis of an infallible criterion.

New Form Equated to Old Form 1

Conventional equating methods (Tucker, Levine equally reliable, and Chained equipercentile) used for equating NEW to OLD1 are not affected by different samples taking OLD2 since these samples do not enter into the equating. Thus, the scaled score means for the conventional methods are identical for conditions involving complete data (A, B, and C), and also identical, but different, for conditions involving missing data (D, E, and F). In contrast, since all test forms are calibrated concurrently, 3PL IRT equating results vary slightly across conditions in which the samples taking the other old form vary.

All equating methods are affected by missing responses in the response strings for both the NEW and OLD1 samples (conditions D vs. A, E vs. B, and F vs. C,), although, for this simulation, Chained equipercentile equating appears less affected than the other methods.

New Form Equated to Old Form 2

These equatings, shown in the right-hand subplot of Figure 2, are the interesting ones -- by design they are most affected by the experimental conditions. As seen in Figure 2 and also in Table 1, the benchmark conditions for all equating methods are different from the benchmark conditions for the equating of NEW to OLD1. The Tucker benchmark conditions are most different -- over one and a half scaled score points; the Levine equally reliable benchmark conditions are least different -- less than a fifth of a scaled score point. Differences for the Chained equipercentile and 3PL IRT benchmark conditions are about the same.

The most striking aspect of these equatings, as was the case for the equatings from Stocking et al. (1988) and Eignor et al. (1990) depicted in Figure 3, is the sensitivity of observed-score equating methods to differences in true sample ability. The introduction of samples of unequal ability, whether in the complete data situation (condition B) or in the missing data condition (condition E) has the largest impact on Tucker equating, and less but substantial impact on Chained equipercentile equating. The remaining two methods, Levine equally reliable and 3PL IRT, seem to be affected to about the same degree.

As in the OLD1 equatings, the introduction of missing data (conditions D vs. A, conditions E vs. B and conditions F vs. C) also impacts the projected means, making them slightly lower for all equating methods.

A particular hypothesis presented by Charles Lewis (personal communication, October 21, 1987) for changes in 3PL IRT equating results across missing data representative (random and unequal) sample conditions

(condition E) and matched sample conditions (condition F), and discussed in Lawrence and Dorans (1990), is demonstrated by the decreases in the projected scaled score means between conditions E and F. Tucker and Levine equally reliable are identical, as they must be, under complete data and missing data matched sample conditions (both models reduce to the direct nonanchor linear equating method in which means and standard deviations are set equal for the new form and old form samples; see Lawrence & Dorans, 1990), and the Chained equipercentile equating is reasonably close to them.

If the benchmark condition (Condition A) is used as a criterion, it seems clear that the 3PL IRT and Levine equally reliable equatings vary least across all experimental conditions. If the Missing Data, Equivalent Samples condition (D) is a more practical criterion, in other missing data conditions (E and F), all equating methods except Tucker come closer to this criterion when representative (i.e., random and unequal) samples are used than when matched samples are used. The matching process appears to improve the Tucker method slightly, while making the other methods much worse.

It is useful to compare the shapes of the plots of means for equating NEW to OLD2 contained in Figures 2 and 3. Although these plots differ somewhat for particular equating methods (i.e., compare the Tucker B to C conditions for the replication to the comparable B to C conditions for the original study and the current study--test variation), in general they are comparable in appearance and the conclusions that may be drawn from all three are the same. In addition, while the introduction of test variation seems to exacerbate slightly the differences in means across conditions for the various equating methods when compared to the situation when a test is equated to

itself, this change in the differences was not as large as anticipated. Based on the results of this single simulation with test variation, it would appear that variations in sample ability and the completeness of response data are greater contributors to differences in means resulting from the various equating methods than are differences in the forms being equated. This conclusion may be partly or wholly due to the fact, however, that forms of the SAT are developed to tight content and statistical specifications, and such results may not have been observed if the simulation were done using data from a test where forms were not so parallel.

The results of this study are essentially the same as the results of the previous studies by Stocking et al. (1988) and Eignor et al. (1990) and suggest that if Levine equally reliable, Chained equipercentile, or 3PL IRT equatings are to be used, more reasonable results are obtained using representative (i.e., random and unequal) samples. If Tucker equating is to be used and there is missing data, better results are obtained with matched samples than with representative but unequal samples. However, if the decision concerning the choice of equating procedure is to be made after the sampling decision, then these results suggest that it is better to use the representative sampling that typically occurs in SAT equating situations, and to avoid selecting the Tucker method.

CONCLUSIONS

As mentioned in the introduction, one criticism of the simulation studies on matching done by Stocking et al. (1988) and Eignor et al. (1990) is that their design called for variations in sample ability and the completeness

of response data while controlling for test variation. Tests were equated to themselves, which does not pattern reality in equating the SAT, where a new form is equated to different old forms. Hence, the results were seen by some as tenuous, because they were not reality-based.

In the current study, test variation was introduced to pattern reality. The results of this study confirm the results of the previous two studies and, collectively, all three studies form a strong foundation for making recommendations about whether to match on a fallible criterion--anchor test score. Only for Tucker equating are better results generally obtained when samples of unequal ability are matched on this fallible criterion.

Caveats presented in the conclusions sections of the previous studies are again relevant. The results of this and the previous studies should be examined from the viewpoint that response data in these simulations were generated according to the 3PL model, with some specific model violations introduced to incorporate missing data. These circumstances may favor the 3PL IRT equating results. Also, it is really not possible to draw definitive conclusions about the importance of the equating differences seen in these studies until estimates of standard errors of equating for all equating methods studied can be produced. However, the very similar patterns of results across the three studies does allow one to conclude, even without the standard errors, that matching samples on anchor test scores is not the best way to proceed in dealing with equating samples of unequal ability.

References

Angoff, W. H. (1971). The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests. New York: College Entrance Examination Board.

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Education Testing Service.

Dorans, N. J. (1990). The equating methods and sampling designs. Applied Measurement in Education, 3, 3-17.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. Applied Measurement in Education, 3, 37-52.

Lawrence, I. M., & Dorans, N. J. (1990). The effect on equating results of matching on an anchor test. Applied Measurement in Education, 3, 19-36.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum, Assoc.

Stocking, M. L., Eignor, D. R., & Cook, L. L. (1988). Factors affecting the sample invariant properties of linear and curvilinear observed- and true-score equating procedures (RR-88-41). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST V user's guide. Princeton, NJ: Educational Testing Service.

Figure 1. Data collection design for equating the SAT

	NEW	EQ1	EQ2	OLD1	OLD2
Sample 1	X	X			
Sample 2	X		X		
Sample 3		X		X	
Sample 4			X		X

Notes: An X denotes the specific total test and anchor test taken by a specific sample.

Samples 1 and 2 are representative samples from the same total group.

Sample 3 is a sample from a different total group that is similar in ability to the total group from which Samples 1 and 2 were drawn.

Sample 4 is a sample from a different total group that is dissimilar in ability to the total group from which Samples 1 and 2 were drawn.

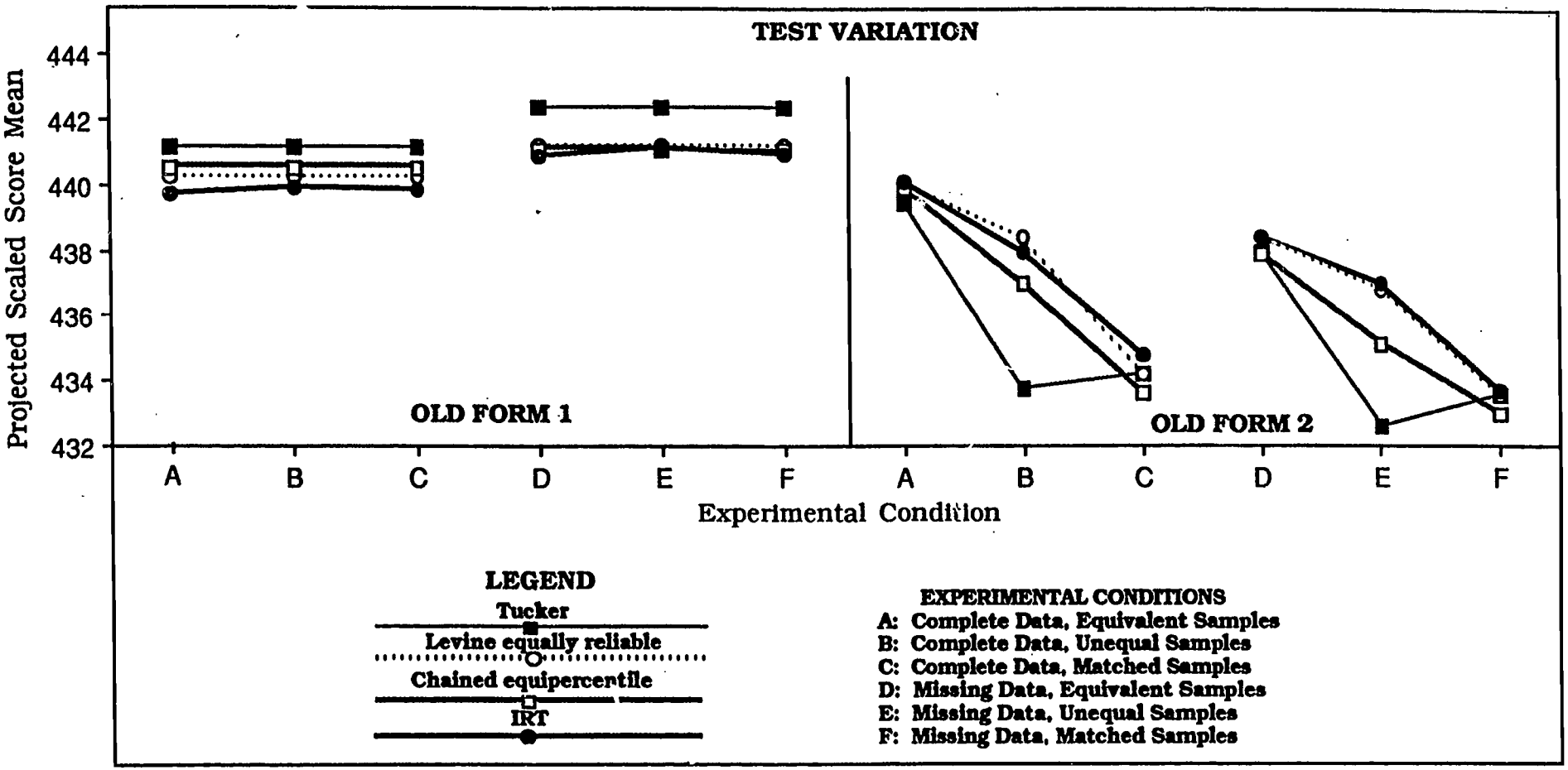

Figure 2. Projected scaled score means for all equating methods and all experimental conditions from test variation phase of study.

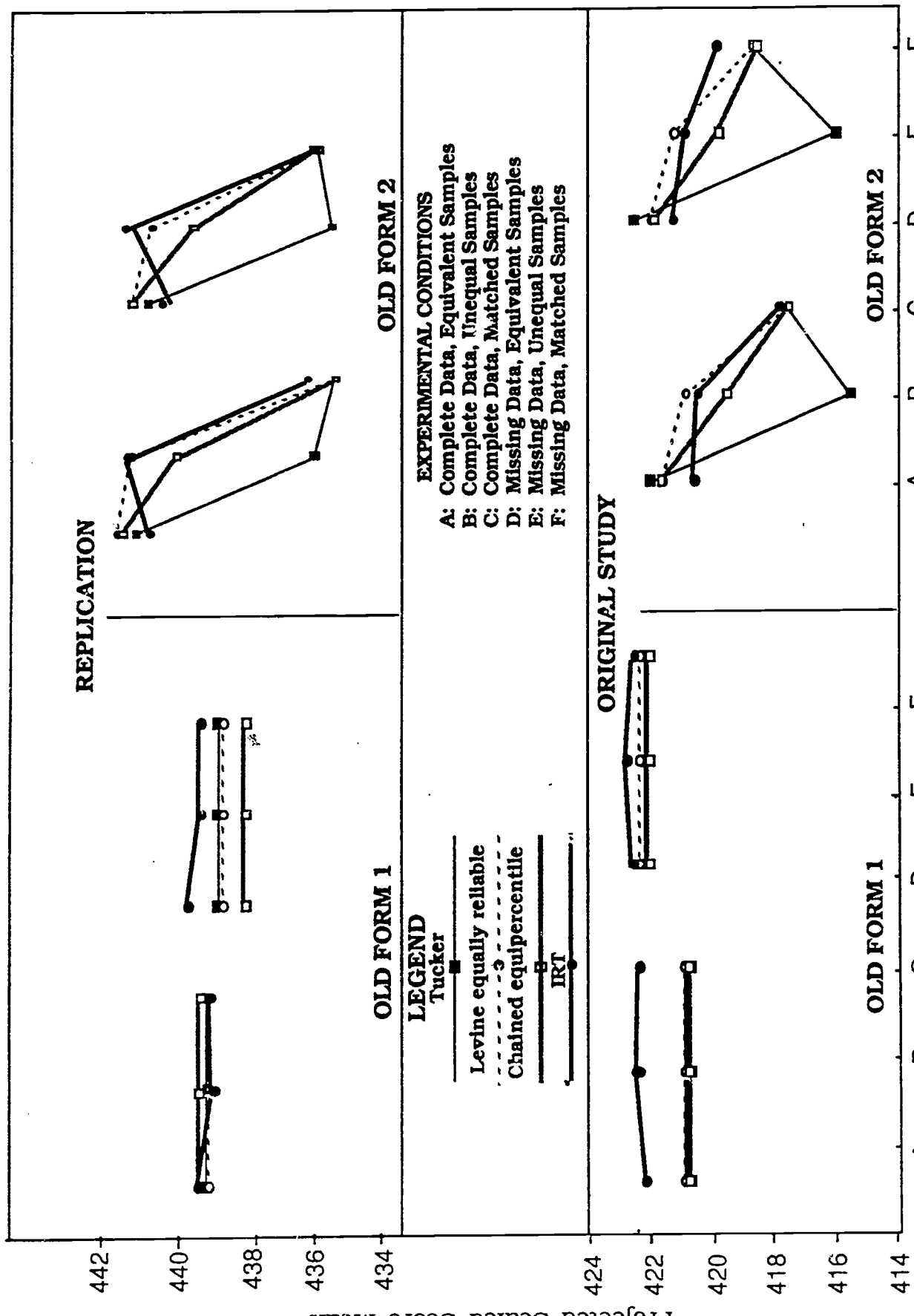

Figure 3. Projected scaled score means for all equating methods and all experimental conditions from original study and replication.

Table 1

Projected Scaled Score Means and Standard Deviations for All Equating Methods and All Experimental Conditions

- Test Variation -

Tucker

	NEW to OLD1		NEW to OLD2		Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Complete Data, Equivalent Samples (Benchmark)	441.14	110.74	439.45	108.02	440.14	110.73
Complete Data, Unequal Samples	441.14	110.74	433.67	104.77	437.24	107.52
Complete Data, Matched Samples	441.14	110.74	434.12	108.83	437.47	109.55
Missing Data, Equivalent Samples	442.37	112.20	437.94	106.31	439.99	109.01
Missing Data, Unequal Samples	442.37	112.20	432.54	102.44	437.29	107.09
Missing Data, Matched Samples	442.37	112.20	433.50	105.65	437.77	108.69

Levine equally reliable

	NEW to OLD1		NEW to OLD2		Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Complete Data, Equivalent Samples (Benchmark)	440.22	110.73	440.05	109.03	440.08	110.79
Complete Data, Unequal Samples	440.22	110.73	438.44	104.49	439.25	107.48
Complete Data, Matched Samples	440.22	110.73	434.12	108.83	437.09	109.66
Missing Data, Equivalent Samples	441.22	112.56	438.43	107.13	439.67	109.62
Missing Data, Unequal Samples	441.22	112.56	436.73	102.04	438.82	107.07
Missing Data, Matched Samples	441.22	112.56	433.50	105.65	437.20	108.87

Chained equipercentile

	NEW to OLD1		NEW to OLD2		Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Complete Data, Equivalent Samples (Benchmark)	440.48	109.96	439.93	108.79	440.17	109.29
Complete Data, Unequal Samples	440.48	109.96	436.91	106.48	438.66	108.13
Complete Data, Matched Samples	440.48	109.96	433.58	108.39	436.99	109.09
Missing Data, Equivalent Samples	441.09	110.51	437.90	106.85	439.38	108.54
Missing Data, Unequal Samples	441.09	110.51	435.07	103.81	437.96	107.01
Missing Data, Matched Samples	441.09	110.51	432.94	104.99	436.90	107.61

IRT

	NEW to OLD1		NEW to OLD2		Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Complete Data, Equivalent Samples (Benchmark)	439.70	107.92	440.13	107.36	439.91	107.63
Complete Data, Unequal Samples	439.92	107.53	437.97	106.45	438.95	106.96
Complete Data, Matched Samples	439.87	107.77	434.71	106.86	437.29	107.29
Missing Data, Equivalent Samples	440.90	107.11	438.45	105.87	439.68	106.46
Missing Data, Unequal Samples	441.16	106.74	436.94	104.93	439.05	105.79
Missing Data, Matched Samples	440.93	106.86	433.65	104.02	437.29	105.40

Table A1

Projected Scaled Score Means and Standard Deviations for All Equating Methods and All Experimental Conditions

- Original Study -

Tucker

	NEW to OLD1		NEW to OLD2		Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Complete Data, Equivalent Samples (Benchmark)	420.72	112.39	421.22	108.52	420.96	110.44
Complete Data, Unequal Samples	420.72	112.39	414.90	106.31	417.80	109.34
Complete Data, Matched Samples	420.72	112.39	416.83	111.09	418.76	111.73
Missing Data, Equivalent Samples	422.10	111.14	421.71	109.14	421.89	110.13
Missing Data, Unequal Samples	422.10	111.14	415.35	107.02	418.71	109.07
Missing Data, Matched Samples	422.10	111.14	417.95	108.92	420.02	110.02

Levine equally reliable

	NEW to OLD1		NEW to OLD2		Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Complete Data, Equivalent Samples (Benchmark)	420.89	112.30	420.79	107.55	420.83	109.91
Complete Data, Unequal Samples	420.89	112.30	420.06	106.97	420.47	109.62
Complete Data, Matched Samples	420.89	112.30	416.83	111.09	418.85	111.68
Missing Data, Equivalent Samples	422.31	110.87	421.15	108.42	421.73	109.63
Missing Data, Unequal Samples	422.31	110.87	420.42	108.01	421.36	109.43
Missing Data, Matched Samples	422.31	110.87	417.95	108.92	420.13	109.88

Chained equipercentile

	NEW to OLD1		NEW to OLD2		Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Complete Data, Equivalent Samples (Benchmark)	420.74	112.77	420.82	107.85	420.81	110.24
Complete Data, Unequal Samples	420.74	112.77	418.76	107.39	419.78	110.00
Complete Data, Matched Samples	420.74	112.77	416.86	111.10	418.84	111.86
Missing Data, Equivalent Samples	422.00	110.67	421.05	108.24	421.52	109.38
Missing Data, Unequal Samples	422.00	110.67	419.04	108.02	420.52	109.28
Missing Data, Matched Samples	422.00	110.67	417.82	108.93	419 90	109.72

IRT

	NEW to OLD1		NEW to OLD2		Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Complete Data, Equivalent Samples (Benchmark)	422.12	111.10	419.79	109.13	420.95	110.12
Complete Data, Unequal Samples	422.35	110.99	419.70	109.56	420.76	110.27
Complete Data, Matched Samples	422.34	111.18	417.11	110.84	419.73	111.01
Missing Data, Equivalent Samples	422.52	110.37	420.46	108.94	421.49	109.65
Missing Data, Unequal Samples	422.77	110.17	420.12	109.90	421.45	110.04
Missing Data, Matched Samples	422.50	110.33	419.07	108.68	420.79	109.50

Table A2

Projected Scaled Score Means and Standard Deviations for All Equating Methods and All Experimental Conditions

- Replication -

Tucker

	NEW to OLD1		NEW to OLD2		Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Complete Data, Equivalent Samples (Benchmark)	439.05	108.92	440.52	106.29	439.78	107.60
Complete Data, Unequal Samples	439.05	108.92	435.20	105.67	437.12	107.29
Complete Data, Matched Samples	439.05	108.92	434.63	107.60	436.84	108.26
Missing Data, Equivalent Samples	438.75	108.09	440.16	106.38	439.46	107.23
Missing Data, Unequal Samples	438.75	108.09	434.70	104.64	436.73	106.37
Missing Data, Matched Samples	438.75	108.09	435.12	107.11	436.94	107.60

Levine equally reliable

	NEW to OLD1		NEW to OLD2		Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Complete Data, Equivalent Samples (Benchmark)	439.02	109.21	441.07	106.03	440.04	107.62
Complete Data, Unequal Samples	439.02	109.21	440.67	106.23	439.84	107.71
Complete Data, Matched Samples	439.02	109.21	434.63	107.60	436.82	108.40
Missing Data, Equivalent Samples	438.60	108.20	440.62	106.22	439.61	107.21
Missing Data, Unequal Samples	438.60	108.20	440.05	105.55	439.32	106.88
Missing Data, Matched Samples	438.60	108.20	435.12	107.11	436.86	107.66

Chained equipercentile

	NEW to OLD1		NEW to OLD2		Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Complete Data, Equivalent Samples (Benchmark)	439.20	109.20	440.92	106.16	440.03	107.57
Complete Data, Unequal Samples	439.20	109.20	439.35	106.16	439.25	107.58
Complete Data, Matched Samples	439.20	109.20	434.63	107.43	436.89	108.22
Missing Data, Equivalent Samples	438.04	107.08	440.61	106.45	439.07	106.40
Missing Data, Unequal Samples	438.04	107.08	438.86	105.79	438.21	106.11
Missing Data, Matched Samples	438.04	107.08	435.22	107.22	436.38	106.80

IRT

	NEW to OLD1		NEW to OLD2		Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Complete Data, Equivalent Samples (Benchmark)	439.26	108.54	440.12	107.36	439.69	107.95
Complete Data, Unequal Samples	438.89	108.31	440.78	108.20	439.83	108.26
Complete Data, Matched Samples	438.95	108.52	435.44	107.73	437.20	108.12
Missing Data, Equivalent Samples	439.58	108.06	439.77	106.91	439.67	107.49
Missing Data, Unequal Samples	439.20	107.85	440.82	107.80	440.01	107.82
Missing Data, Matched Samples	439.24	108.03	435.26	107.64	437.25	107.84